Transferencia de Aprendizaje para Clasificación de Peatones

Genaro Camele[1], Facundo Quiroga[1], Franco Ronchetti[1], Waldo Hasperué[1,2], and Laura Lanzarini[1]

[1] Instituto de Investigación en Informática LIDI, Facultad de Informática, Universidad Nacional de La Plata,
[2] Investigador Asociado - Comisión de Investigaciones Científicas (CIC)
{gcamele,fquiroga,fronchetti,whasperue,laural}@lidi.info.unlp.edu.ar

Resumen Desde la aparición, en el año 2005, de los Histogramas de Gradientes Orientados (*HOG*) como descriptor para detección de peatones, han aparecido numerosas publicaciones que permitieron mejorar la clasificación agregando mejoras o comparando con nuevos descriptores. Del mismo modo, año a año aparecen nuevas bases de datos con imágenes de entornos reales, que permiten la evaluación de los modelos desarrollados. No obstante, la utilización de un modelo entrenado en un entorno real nuevo no siempre resulta trivial y es un tema pendiente de estudio para este dominio.
En este artículo, presentamos un protocolo para evaluar la transferencia de aprendizaje entre tres de las bases de datos más utilizadas en la literatura: *INRIA, Daimler* y *TUD-Brussels*. Comparamos los descriptores HOG y Patrones Binarios Locales (LBP) en conjunto con Máquinas de Vectores de Soporte (SVM) como clasificador de base. Los resultados obtenidos muestran que si bien cada conjunto de datos presenta escenas del mundo real, existen diferencias significativas que hacen que un modelo entrenado con un conjunto de imágenes no funcione apropiadamente con otro. Por otro lado, encontramos que al entrenar un modelo con la mezcla de diferentes bases de datos permite una mayor transferencia de aprendizaje, si bien no siempre ayuda al entrenamiento de un conjunto de datos particular.

Keywords: detección de peatones, transferencia de aprendizaje, SVM, HOG, LBP, Daimler, Inria, TUD-Brussels

1. Introducción y estado del arte

La detección de peatones puede verse como uno de los temas más importantes de la detección de objetos dentro del área de visión por computador. En los últimos años ha atraído la atención de muchos investigadores, lo que resultó en grandes avances debido a su impacto social y aplicaciones en seguridad vial [12][1]. En los últimos años, ya superada la barrera de la detección con una tasa de acierto relativamente alta, los investigadores comenzaron a enfocarse en

problemas puntuales como la detección de personas parcialmente ocultas, o la detección de imágenes borrosas o con poca resolución[8].

La detección de peatones, como otras tareas de detección de objetos, consiste principalmente de tres partes: el cálculo de un descriptor de imágenes adecuado para representar a un peatón, un modelo de clasificación que distinga imágenes de peatones (positivas) de las que no son (negativas) en base al descriptor, y un sistema de ventana deslizante que encuentra los peatones en una escena realizando varias ejecuciones del clasificador. En este artículo nos enfocamos en las primeras dos etapas ya que son las más importantes para evaluar la transferencia de aprendizaje.

Existen numerosas aproximaciones para llevar a cabo esta tarea en cuanto al cómputo de descriptores y métodos de clasificación. Si bien existen diversos artículos que investigan descriptores como Haar, detección de bordes, o histogramas de color, el enfoque más ampliamente estudiado, y que mejores resultados ha dado, es la combinación de descriptores HOGs (*Histogram of Oriented Gradients* con SVM [3][4]. En el mismo sentido, podemos encontrar en segundo lugar a los descriptores LBP (*Local binary pattern*) generalmente utilizados en combinación con los descriptores HOGs [10][9][6].

Una suposición importante en muchos enfoques tradicionales de detección de peatones es que el conjunto de datos de entrenamiento y el conjunto de datos de prueba deben ser similares. Esto se deduce del hecho de que las imágenes de entrenamiento y las de prueba son tomadas de la misma base de datos. Esta premisa dificulta transferir un modelo entrenado a un entorno real, donde las condiciones no son iguales a las que había en la base de datos original. Es decir, las características de las imágenes de las distintas bases de datos siguen una distribución diferente [1].

Hay varias razones por las cuales las características de distintas bases de datos varian. Por un lado, puede diferir el proceso de captura, por ejemplo debido a variaciones climatológicas que inciden en la luz o la temperatura de color. En el mismo sentido, las especificaciones técnicas de la cámara de captura podrían no ser similares. Por otro lado, el contexto en que se captura la base de datos nunca es el mismo, presentando problemas aún mayores como las variaciones en la vestimenta de los peatones y el fondo de la escena. Por ejemplo, algunas bases de datos presentan escenas con naturaleza de fondo y otras se capturaron en contextos de tráfico en ciudades o autopistas [3][5][11]. Si bien la utilización de un descriptor apropiado ayuda a que exista una relación más cercana entre el espacio de las imágenes de entrenamiento con respecto a las de prueba, no son lo suficientemente invariantes a estas diferencias en las características de las bases de datos.

En este contexto, existen muy pocos trabajos que utilicen diversas bases de datos existentes para evaluar la posibilidad de utilizar un modelo entrenado en escenas no vistas. En [2] realizan un proceso de transferencia con el objetivo de generar un modelo más robusto a variaciones en las imágenes, pero se enfocan en aumentar el desempeño en una base de datos particular y no en evaluar la capacidad de transferir el aprendizaje entre bases de datos. En [1] los autores

analizan la transferencia de aprendizaje para algunas bases de datos. No obstante, al utilizar la tarea de detección como medida de desempeño en lugar de la de clasificación introducen factores de variabilidad mayores que dificultan la interpretación de la transferencia de aprendizaje.

En este trabajo nos proponemos evaluar la transferencia de aprendizaje de modelos de *clasificación* de peatones en tres bases de datos ampliamente utilizadas: INRIA [3], Daimler [5] y TUD-Brussels [11]. Para ello establecimos un protocolo de experimentación para generar lo que definimos como *matrices de transferencia*. Las mismas reflejan, aproximadamente, la similitud entre las distintas bases de datos, así como la complementariedad de las mismas para generar un modelo de clasificación robusto y capaz de generalizar a nuevas escenas. De esta manera, podemos evaluar posibilidad de transferir un modelo entrenado con una base de datos a nuevas escenas con características de imagen diferentes.

2. Bases de datos y descriptores

A continuación presentamos las tres bases de datos que comparamos en los experimentos, Daimler Mono Pedestrian Detection, INRIA Person y TUD-Brussels Motionpairs. Además, se introducen brevemente los dos descriptores analizados, elegidos en base a los más utilizados en el estado del arte.

2.1. Bases de datos

La base de datos **INRIA Person** [3] (Figura 1) contiene imágenes de personas en varias situaciones. Fue generada en base a varios corpus e imágenes recolectadas de forma independiente. Las mismas son a color y de tamaño 64×128 píxeles. El conjunto de entrenamiento contiene 3.030 y 34.892 ejemplos positivos y negativos respectivamente, y el de prueba 1.028 y 453.

Figura 1: Ejemplos positivos (izquierda) y negativos (derecha) de la base de datos INRIA Person[3].

La base de datos **Daimler Mono Pedestrian Detection** [5] (Figura 2) se enfoca en la detección de peatones en cámaras monoculares. Contiene imágenes en escala de grises tomadas desde un auto en movimiento. El conjunto de

entrenamiento contiene 15.560 imágenes de peatones de 48×96 píxeles. Además, se incluyen 6.744 imágenes originales (tamaño completo) en donde no hay peatones con el objetivo de ser utilizadas para extraer patrones negativas. Con estas imágenes se generaron 32.000 patrones negativos del mismo tamaño que las imágenes positivas tomando muestras de forma aleatoria. Luego, dividimos los conjuntos de imágenes utilizando el 90 % para entrenamiento y el 10 % para evaluación.

Figura 2: Ejemplos positivos (arriba) y negativos (abajo) de la base de datos Daimler Mono Pedestrian Detection[5].

La base de datos **TUD-Brussels Motionpairs** [11] (Figura 3) también se enfoca en la detección de peatones desde un auto en movimiento, con imágenes a color de 640×480 píxeles. Contiene 551 peatones etiquetados para el conjunto de entrenamiento y 311 para el de prueba. El cuadro 1 resume la información básica de las tres bases de datos utilizadas.

Figura 3: Ejemplos de la base de datos TUD-Brussels Motionpairs [11].

2.2. Descriptores

El **Histograma de Gradientes Orientados (HOGs)** [3] se calcula en base al gradiente de la imagen. Dado que dicho gradiente tiende a ser más grande en los bordes de los objetos, es apto para representar la forma de los mismos. En

Cuadro 1: Detalle de los conjuntos de datos utilizados.

Base	Entrenamiento		Prueba		Color	Contexto
de datos	Positivas	Negativas	Positivas	Negativas		
Daimler	14.000	30.000	1.560	2.000	No	Peatones
INRIA	3.030	34.879	1.028	453	Sí	Personas
TUD-Brussels	982	9.064	110	1.010	Sí	Peatones

base al gradiente, representado como ángulo-magnitud, se divide la imagen en celdas de igual tamaño. Para cada celda se calcula un histograma con n cubetas. Cada cubeta corresponde a un rango de ángulos, y cada punto del gradiente contribuye a la cubeta en forma proporcional a su magnitud. Luego los histogramas de las celdas contiguas se agrupan en bloques, típicamente superpuestos, que se normalizan. Los bloques normalizados luego pierden su estructura 2D y se concatenan en un vector unidimensional de modo de formar el descriptor final. La última fila de la Figura 4 presenta una visualización de las celdas del HOG de imágenes positivas y negativas de cada dataset.

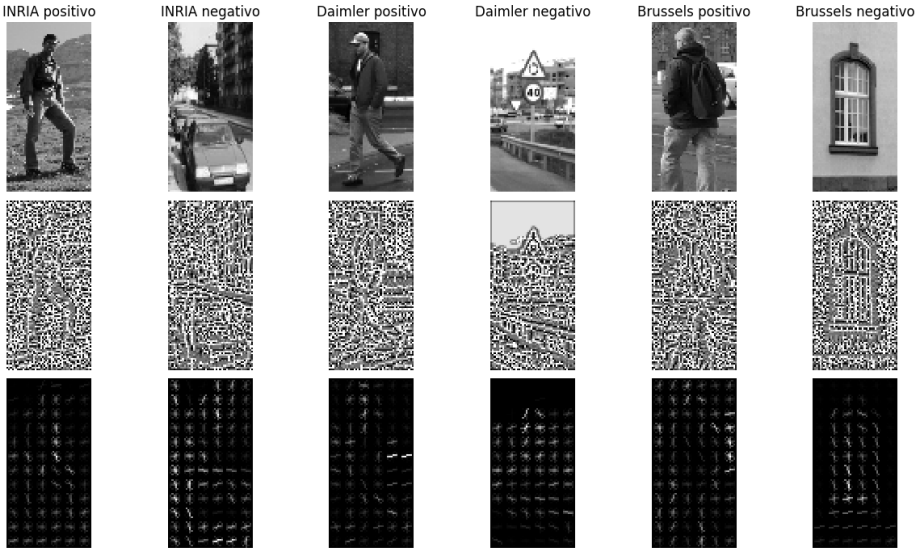

Figura 4: Imágenes positivas y negativas originales de cada una de las bases de datos (fila 1). LBPs y HOGs calculados para cada imagen (filas 2 y 3).

Patrones Binarios Locales (LBP). Los **Patrones Binarios Locales (LBP)** [7] son descriptores de textura. Se calculan dividiendo la imagen en celdas, y por

cada pixel de la celda se calcula un número binario de n dígitos que representa el gradiente en un camino de la vecindad del pixel, donde n es el tamaño del camino. Luego, se genera un histograma con 2^n cubetas para toda la celda, donde cada pixel incrementa en 1 la cubeta correspondiente a ese número binario. Al igual que con los HOGs los histogramas se concatenan para formar el descriptor final. La segunda fila de la Figura 4 también presenta una visualización del descriptor.

3. Experimentos

3.1. Protocolo del experimento

A continuación presentamos el protocolo de experimentación para generar las matrices de transferencia de aprendizaje. En primer lugar, entrenamos un modelo distinto para cada una de las tres bases de datos, utilizando su conjunto de entrenamiento respectivo. Luego, por cada modelo, se realiza la evaluación del mismo con los tres conjuntos de evaluación de las bases de datos. Como métrica de desempeño se utilizó la medida-F (*F-Score*), de modo de combinar la precisión (*precision*) y la exhaustividad (*recall*) en una sola métrica.

De este modo, obtuvimos una matriz de transferencia D, de tamaño 3×3, entre las tres bases de datos, donde $D_{i,j}$ nos indica la medida-F del modelo al ser entrenado con la base de datos i y evaluado con la base de datos j. De este modo, podemos observar, la similitud que hay entre las características de las imágenes de las diferentes bases de datos, para cada descriptor.

3.2. Preprocesamiento de los datos y entrenamiento del modelo

Dado que no todas las bases de datos tienen el mismo tamaño de imagen, unificamos el mismo a 48×96 píxeles como medida estándar para llevar a cabo los experimentos. Esto no supone ninguna deformación de relación de aspecto debido a que en todos los casos es de $1:2$.

Convertimos las imágenes a color de INRIA y TUD-Brussels a escala de grises de modo que el cálculo del HOG y los LBPs sea el mismo para los distintos conjuntos de datos. Normalizamos el rango de las imágenes al intervalo $[0 \dots 1]$. No realizamos una ecualización del histograma ni ningún otro tipo de preprocesamiento adicional a las imágenes para preservar las distribuciones originales de los datos.

Calculamos los HOGs sobre la imagen resultante con un tamaño de celda de 8×8 píxeles, con bloques de 2×2 celdas. Aplicamos normalización L2 a los descriptores obtenidos, siguiendo la metodología de [3]. Los LBPs se calculan con 8 puntos de muestreo en un radio de 1 pixel, siguiendo también la metodología de [10].

Como modelo de clasificación utilizamos Máquinas de Vector de Soporte ya que es el modelo más utilizado en la literatura al mismo tiempo de ser simple y muy eficiente computacionalmente para luego incorporarlo a un esquema de detección. En este caso utilizamos un SVM con una configuración similar a [3]

donde $C = 0{,}1$, el núcleo es lineal, igual peso para las dos clases y regularización L2. Entrenamos el modelo mediante el método SMO de la librería *liblinear* con la formulación dual, verificando que el modelo converge y no sobreajusta.

3.3. Resultados y discusión

El cuadro 2 muestra las matrices de transferencia entre los tres conjuntos de datos, para cada uno de los descriptores utilizados. Claramente, la transferencia de aprendizaje no es trivial. Esto queda reflejado al ver que la diagonal principal de cada matriz, que contiene los resultados de entrenar y evaluar un modelocon la misma base de datos, tiene una medida-F muy superior al resto de las entradas. Si bien Daimler parece tener mejor capacidad de transferencia, es posible que se deba a la mayor cantidad de ejemplos de esa base de datos.

Por otro lado, podemos observar que el mejor resultado obtenido de manera general es utilizando el descriptor HOG. El agregado del descriptor LBP como parte representativa de una imagen no parece generar gran impacto sobre la precisión del modelo, considerando el aumento significativo en la dimensión del descriptor final. No obstante, este agregado demuestra mejorar significativamente la transferencia de aprendizaje para algunos casos como por ejemplo al entrenar con la base de datos INRIA y evaluar con las otras dos. Esto indica que la transferencia es muy dependiente del descriptor utilizado. Es posible que esto se deba también al hecho de que INRIA es una base de datos de personas y no de peatones exclusivamente.

Cuadro 2: Matrices de transferencia entre las tres bases de datos utilizadas. Las entradas muestran los *F-Score*. Filas: Entrenamiento. Columnas: Evaluación. Abreviaciones: I = INRIA, D = Daimler, B = TUD-Brussels.

(a) HOG

	I	D	B
I	0.619	0.297	0.064
D	0.841	0.984	0.627
B	0.335	0.396	0.803

(b) LBP

	I	D	B
I	0.176	0.078	0.097
D	0.693	0.887	0.278
B	0.167	0.151	0.218

(c) HOG+LBP

	I	D	B
I	0.611	0.351	0.210
D	0.837	0.979	0.556
B	0.376	0.369	0.748

3.4. Transferencia al entrenar con múltiples bases de datos

Una extensión natural del experimento anterior consiste en unir los conjuntos de entrenamiento de dos bases de datos para generar el modelo, y realizar la evaluación de igual forma que en el experimento anterior. De esta forma podemos evaluar si el agregado de datos diversos contribuye a la mejora de la transferencia. Es decir, analiza la complementariedad de las bases de datos.

El Cuadro 3 muestra las matrices de transferencia resultantes. Nuevamente, la métrica utilizada es la medida-F.

Cuadro 3: Matrices de transferencia entre las tres bases de datos utilizadas al entrenar con múltiples bases de datos. Las entradas muestran los *F-Score*. Filas: Entrenamiento. Columnas: Evaluación. Abreviaciones: I = INRIA, D = Daimler, B = TUD-Brussels.

(a) HOG

	I	D	B
I+D	0.837	0.960	0.578
I+B	0.747	0.645	0.658
D+B	0.816	0.976	0.850

(b) LBP

	I	D	B
I+D	0.444	0.709	0.215
I+B	0.273	0.149	0.176
D+B	0.581	0.829	0.303

(c) HOG+LBP

	I	D	B
I+D	0.844	0.946	0.472
I+B	0.697	0.562	0.497
D+B	0.789	0.965	0.766

Podemos observar que al entrenar con varias bases de datos aumenta la medida-F en casi todos los casos con respecto a usar una sola base de datos. No obstante, debemos considerar un efecto de tamaño por la unión de los conjuntos de entrenamiento. Por ejemplo, TUD-Brussels no aumenta de forma considerable los ejemplos al sumarlos a los de INRIA y sin embargo aumenta significativamente la medida-F de su conjunto de evaluación. La figura 5 resume los resultados obtenidos en todos los experimentos utilizando el descriptor HOG, que fue el que mejor desempeño logró. Las diferentes series muestran las distintas configuraciones de entrenamiento, mientras que las 3 columnas principales representan las tres bases de datos de evaluación.

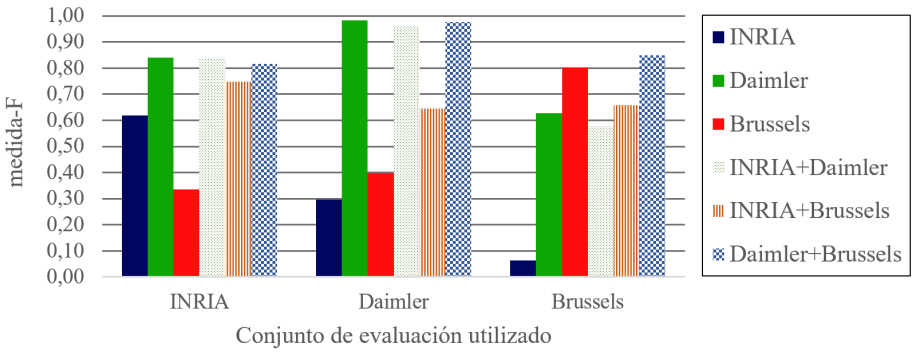

Figura 5: Gráfico comparativo de la medida-F para la evaluación de cada base de datos al entrenar con todas las combinaciones de bases de datos analizadas, utilizando el descriptor HOG.

4. Conclusión

En este artículo presentamos un protocolo para evaluar la transferencia de aprendizaje entre bases de datos de peatones. Los resultados indican que si bien los conjuntos de datos más ampliamente utilizados para la detección de peatones tienen la misma finalidad, presentan diferencias singificativas que hacen no trivial la utilización de un modelo entrenado en un entorno real, con cambios en la luminosidad o condiciones de escena. Ninguna de las tres bases de datos resulta ideal como único modelo de entrenamiento para llevar a cabo un proceso de transferencia de aprendizaje, aunque, en base a los resultados Daimler, parece la más apta de las evaluadas.

En trabajos futuros, buscaremos ampliar el rango de bases de datos para aumentar la riqueza de la comparación, así como cuantificar el efecto del tamaño de las mismas. Además, esperamos evaluar modelos más recientes del estado del arte como aquellos basados en Redes Convolucionales. Una vez terminada la evaluación de la transferencia para clasificación, proponemos ampliar los experimentos para incluir tareas de detección con el fin de evaluar su indicidencia.

Referencias

1. Benenson, R., Omran, M., Hosang, J., Schiele, B.: Ten years of pedestrian detection, what have we learned? In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) Computer Vision - ECCV 2014 Workshops. pp. 613–627. Springer International Publishing, Cham (2015)
2. Cao, X., Wang, Z., Yan, P., Li, X.: Transfer learning for pedestrian detection. Neurocomputing 100, 51–57 (2013), http://www.sciencedirect.com/science/article/pii/S0925231212003256, special issue: Behaviours in video
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 1, pp. 886–893. IEEE (2005)
4. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. 34(4), 743–761 (Apr 2012), http://dx.doi.org/10.1109/TPAMI.2011.155
5. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection: Survey and experiments. IEEE Transactions on Pattern Analysis & Machine Intelligence (12), 2179–2195 (2008)
6. Gan, G., Cheng, J.: Pedestrian detection based on hog-lbp feature. 2011 Seventh International Conference on Computational Intelligence and Security pp. 1184–1187 (2011)
7. Mu, Y., Yan, S., Liu, Y., Huang, T., Zhou, B.: Discriminative local binary patterns for human detection in personal album. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. pp. 1–8. IEEE (2008)
8. Ouyang, W., Wang, X.: Single-pedestrian detection aided by multi-pedestrian detection. 2013 IEEE Conference on Computer Vision and Pattern Recognition pp. 3198–3205 (2013)
9. Pei, W.J., Zhang, Y.L., Zhang, Y., Zheng, C.H.: Pedestrian detection based on hog and lbp. In: Huang, D.S., Bevilacqua, V., Premaratne, P. (eds.) Intelligent Computing Theory. pp. 715–720. Springer International Publishing, Cham (2014)

10. Wang, X., Han, T.X., Yan, S.: An hog-lbp human detector with partial occlusion handling. In: Computer Vision, 2009 IEEE 12th International Conference on. pp. 32–39. IEEE (2009)
11. Wojek, C., Walk, S., Schiele, B.: Multi-cue onboard pedestrian detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). vol. 00, pp. 794–801 (06 2009), doi.ieeecomputersociety.org/10.1109/CVPRW.2009.5206638
12. Yan, J., Zhang, X., Lei, Z., Liao, S., Li, S.Z.: Robust multi-resolution pedestrian detection in traffic scenes. 2013 IEEE Conference on Computer Vision and Pattern Recognition pp. 3033–3040 (2013)